THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 61, No. 9, November 1982
Printed in U.S.A.

An Improved Model for Isolated Word Recognition

By J. M. TRIBOLET,* L. R. RABINER, and J. G. WILPON

(Manuscript received April 19, 1982)

Current models for isolated word recognition perform very well on small vocabularies of distinctly different sounding words. However, when we are confronted with vocabularies of similar sounding words (e.g., the letters of the alphabet), the performance of isolated word recognizers decreases dramatically. By carefully reexamining the model used for isolated word recognition we have identified some of the inherent deficiencies. In this paper we propose an improved word-recognition model that is inherently capable of accurately recognizing words from almost any vocabulary. We have investigated a simple implementation of the model that preserves most of the structure of a linear predictive coding (LPC)-based version of the canonic isolated word model. In an experimental evaluation of the improved model, using an alpha-digit vocabulary, recognition accuracy improvements of from 1 to 5.7 percent were obtained for four talkers. The improvements were due to changes in both the analysis model and the decision procedure. The strengths and weaknesses of the improved model are discussed.

I. INTRODUCTION

Although the goal of continuous speech recognition by machine remains far out of reach, the one area of speech recognition that is practical with today's technology and understanding is that of isolated word recognition.[1-6] What is interesting about this area is that the general approach used to solve the isolated word-recognition problem (i.e., the statistical-pattern-recognition approach) bears little relationship to the way in which humans understand speech. As a result the vocabularies for which the isolated word recognizers can achieve good

* Work performed while a consultant to Bell Laboratories.

performance are severely constrained in both size and complexity.[7] If we are interested in using a vocabulary for which the performance of the isolated word recognizer is less than perfect (e.g., the letters of the alphabet), then we have to rely on the syntax and semantics of the recognition task to provide the desired level of performance from the overall system.[8-10]

In an effort to improve word-recognition accuracy for arbitrary vocabularies, we have re-examined the word-recognition model and proposed a somewhat more general structure. The proposed changes in the model include an improved feature analysis in which both long-time and short-time features are measured, and an improved decision box in which the two-pass decision rule of Rabiner and Wilpon[11] is adapted to the speaker-trained case.

The implementation of the improved word-recognition model, which we have studied, is based on the standard linear predictive coding (LPC) word recognizer as originally proposed by Itakura.[12] In an effort to retain as much of the original structure as possible, we have used the standard LPC analysis as the long-time features, and a new LPC analysis based on 15-ms frames as the short-time features. Experimentation with the improved model, using a 39-word vocabulary of the alphabet, the digits, and three command words in a speaker-trained mode, showed improvements in accuracy of from 1 to 5.7 percent for four talkers. An analysis of the results showed that the improved feature analysis provided only small improvements in accuracy (from 0 to 1.3 percent), whereas the two-pass decision rule provided somewhat larger improvements in accuracy (from 1 to 4.4 percent).

The outline of this paper is as follows. In Section II we briefly review the canonic isolated word-recognition model and discuss its strengths and weaknesses. We also discuss, in this section, the implementation of the model based on LPC feature analysis and an LPC distance measure. In Section III we present the improved word-recognition model and discuss how it was implemented within the structure of the LPC-based recognizer. In Section IV we describe the experimental evaluation of the improved model based on the alpha-digit vocabulary. Finally, in Section V we discuss the results and their implications for practical systems.

II. THE CANONIC MODEL FOR ISOLATED WORD RECOGNITION

Figure 1 is a block diagram of the canonic (statistical-pattern-recognition) model for isolated word recognition. The three basic components of the model include:

(*i*) Feature measurement in which the speech signal is analyzed to provide a set of Q features (e.g., filter bank energies, LPC coefficients, etc.) once every M samples. If the isolated word is of duration $L \times M$

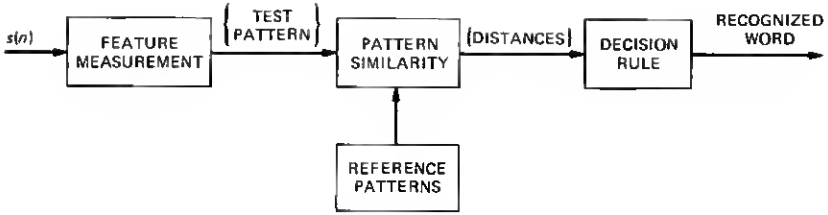


Fig. 1—Block diagram of standard isolated word-recognition model.

samples, then a total of L sets of features characterize the word. The matrix of $Q \times L$ features is called the test pattern.

(*ii*) Pattern similarity measurement in which a score (similarity or distance) relating the similarity of the test pattern to each of a set of V reference patterns is computed. Pattern similarity involves both time alignment (registration) of the test and reference pattern, and distance computation along the alignment path. The output of the pattern similarity box is a set of V distance scores, i.e., one for each reference pattern.

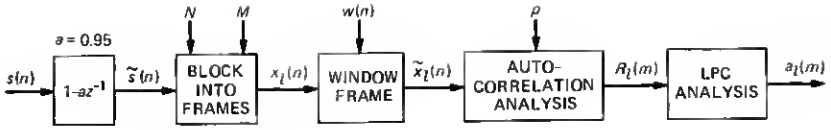
(*iii*) A decision rule in which the distance scores are used to provide an ordered (by distance) list of recognition candidates. Generally, the candidate with the smallest distance is chosen as the "recognized word."

Rather than dwelling further on the canonic model we now review the LPC implementation of this model, as we will be relying on this structure throughout this paper. We will return to the canonic model in Section 2.2 when we discuss its limitations and propose the improved model.

2.1 The LPC-based implementation of the word recognizer

Figure 2 is a block diagram of the feature measurement for an LPC-based analyzer. The digitized speech signal (digitized at a 6.67-kHz rate) is first preemphasized using a simple first-order digital network and then blocked into overlapping frames of N (300) samples with consecutive frames overlapping by 200 samples. Thus, a frame spacing of $M = 100$ samples is used (i.e., 67 frames/second). Each speech frame is then windowed by a 300-sample Hamming window, and a pth-order ($p = 8$) autocorrelation analysis is performed. A full LPC analysis (using the autocorrelation method[13]) is then performed giving the set of $(p + 1)$ LPC coefficients as the features for each frame.

The pattern similarity processing is carried out using a dynamic time-warping (DTW) algorithm in which the test pattern is simultaneously time aligned with each reference pattern, and a distance along the time-alignment path is computed. One of the major features of this



$$\tilde{s}(n) = s(n) - as(n-1)$$

$$x_l(n) = \tilde{s}(Ml + n), \quad l = 0, 1, \ldots, L-1$$
$$n = 0, 1, \ldots, N-1$$

$$\tilde{x}_l(n) = x_l(n) \cdot w(n)$$

$$R_l(m) = \sum_{n=0}^{N-1-m} \tilde{x}_l(n) \quad \tilde{x}_l(n+m)$$

Fig. 2—Block diagram of LPC analysis system.

processing is the local distance measure, which relates the distance between a frame of the test pattern and a frame of the reference pattern, of the form[12]

$$d(T, R) = \log \left[\frac{\mathbf{a}_R \mathbf{V}_T \mathbf{a}_R^t}{\mathbf{a}_T \mathbf{V}_T \mathbf{a}_T^t} \right], \tag{1}$$

where $\mathbf{a}_R$ and $\mathbf{a}_T$ are the LPC feature sets of reference and test, respectively, and $\mathbf{V}_T$ is the autocorrelation coefficient set of the test. The distance measure of eq. (1), called the LPC log-likelihood measure, can be computed using only $(p + 1)$ multiplications and additions, and one logarithm.[12] Furthermore, the LPC distance of eq. (1) has been shown to have reliable and well understood statistical properties.[14,15] In particular, if both $\mathbf{a}_R$ and $\mathbf{a}_T$ are derived from the same underlying stationary random process, then $d(T, R)$ is precisely χ^2 distributed with p degrees of freedom. This statistical behavior of the LPC distance holds for fricative sounds. For voiced speech, although the model is inexact on a frame-by-frame basis, the statistical properties are approximately correct on a time-average basis.

To compute the pattern similarity between the test and each reference pattern using the DTW algorithm with the distance measure of eq. (1), a solution to the minimization of

$$D^* = \min_{w(n)} \left[\sum_{n=1}^{NT} d(T_n, R_{w(n)}) \right] \tag{2}$$

must be found where NT is the number of frames in the test, and $w(n)$ is the warping function relating frame n of the test to frame $w(n)$ of the reference. Efficient recursive procedures for solving eq. (2) have been described in the literature.[12,16–18]

Finally, the decision box orders the distance scores for each reference pattern and chooses either the reference with the minimum distance

(the nearest neighbor rule) or the reference whose average of the K-best scores (for multiple-template systems) is minimum (the K-nearest neighbor rule) as the recognized word. When the recognized word is unique (i.e., only a single reference gets a small distance score), this simple decision rule is sufficient. However, for complex vocabularies, generally several references achieve small distance scores, and reliable recognition using the smallest distance cannot be achieved. In such cases a two-pass decision rule[11] has been shown to increase accuracy by deferring the final recognition decision to a discrimination analysis in a second pass of the decision rule. This discrimination analysis has only been applied to speaker-independent systems because of the problems associated with obtaining appropriate word discrimination weights.[11]

2.2 *Strengths and limitations of the word-recognition model*

The strengths of the canonic word-recognition model of Fig. 1 are as follows:

(*i*) It is invariant to different speech vocabularies, users, feature sets, pattern similarity algorithms, and decision rules.

(*ii*) It is easy to implement.

(*iii*) It works well in practice.

The weaknesses of the model include:

(*i*) The feature analysis only adequately represents long-time stationary events in the speech signal; nonstationary and transient events are only poorly represented.

(*ii*) The model does not perform well for complex vocabularies with acoustically similar words.

We now consider the first weakness of the model. By way of example Fig. 3 shows waveform plots of the beginning regions of two distinct words. Word 1 shows a silence followed by the onset of voiced speech. Word 2 shows a short (15 ms) transient of low-level, unvoiced speech (e.g., a plosive sound) followed by the onset of voiced speech. Figure 3 also shows the placement of the first two long-time speech segments (frames), which contain identical data except for the first 15 ms of the first segment, in which one frame has silence and one frame has a short plosive. It should be clear that for a long-time analysis such as the LPC model of Section 2.1, the low-level differences in the first 15 ms of frame 1 will be swamped out by the high-level voiced speech in the last 30 ms of the frame. Thus, in a long-time stationary framework accurate recognition of differences between short transients and other nonstationary regions (e.g., as occur during onsets and offsets of voicing) is greatly limited. Thus, to ameliorate this weakness, the feature-detection algorithm must be enhanced to include some representation of short-time nonstationary events.

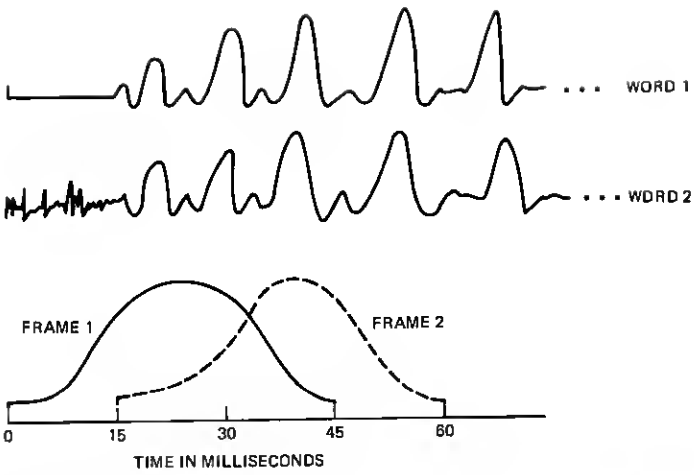


Fig. 3—How a short transient in a word can be swamped out by a voiced region in the long-time analysis model.

Consider now the second weakness of the model. The reason that acoustically similar words are easily confused is that the pattern-similarity measure (the DTW distance) gives equal weight to all frames of the word. For differentiating words of one equivalence class from words of another equivalence class this procedure is reasonable. However, within a class of acoustically similar words a discrimination analysis rather than a straight recognition is required. Such an analysis has been proposed by Rabiner and Wilpon[11] for the case of speaker-independent recognition of words.

For speaker-trained recognizers this two-pass decision rule must be modified so that the optimal weighting curves for word discrimination could be obtained directly from the robust training procedure.[19]

With the incorporation of the expanded feature analysis, a modified DTW algorithm, and an expanded decision rule, the basic weaknesses of the canonic word recognizer can be overcome to some extent. In the next section we describe an "improved" model for word recognition and show how the improvements can be incorporated directly into the LPC framework of Section 2.1.

III. THE IMPROVED WORD-RECOGNITION MODEL

Based on the discussion of Section 2.2, the improved word-recognition model would have a structure of the type shown in Fig. 4. The major differences in the model, from that of Fig. 1, are:

(*i*) The feature measurement box is expanded into three sub-blocks, namely long-time feature measurements, short-time feature

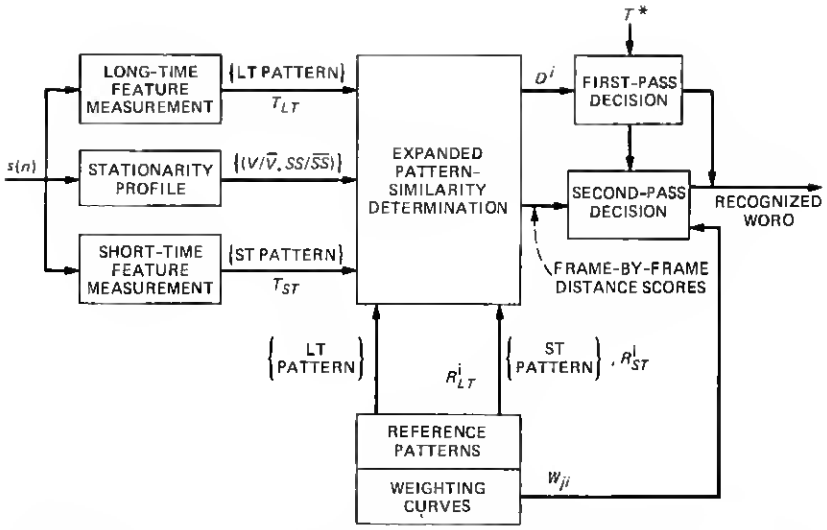


Fig. 4—Block diagram of the improved, isolated word-recognition model using both long-time and short-time features, and a two-pass discrimination model.

measurements, and a stationarity profile. The long-time features are essentially those of the original model, although the rate at which they are measured will generally be higher for this new model than for the original model. The short-time features are intended to characterize transients and other nonstationary events in the speech signal. Some typical short-time features include zero or level crossing counts over short-time intervals, wideband (short-impulse response), filter bank analyses, short-time LPC analyses, etc. The stationarity profile decides which feature set (either long-time or short-time) is used to characterize a given frame of speech, and hence is used for the distance measure of the pattern-similarity algorithm.

(*ii*) The DTW algorithm is expanded to use both long-time and short-time patterns, for both test and reference pattterns, in determining similarity of a given reference pattern to the test pattern. The stationarity profile is used to guide the alignment and to choose which feature set is used in making a given distance computation.

(*iii*) The decision box is implemented as a two-pass decision. In the first-pass decision the distance scores for each reference pattern are ordered, and if the best distance is smaller than the second best distance by a threshold T^*, the decision phase is terminated. If, however, the top two or more references are within T^* in distance, a second-pass decision rule is used in which the similar words are compared using a discriminant analysis and the recognized word is chosen on the basis of this analysis. To implement the discriminant

analysis, a set of distance-weighting curves discriminating word i from word j (for all i, j) must be saved along with the reference patterns. We now describe how the improved model was implemented in the framework of the LPC analysis system.

3.1 The LPC basic improved word-recognition model

Using the LPC analysis framework, the expanded feature measurement was implemented as follows. The long-time analysis was implemented as described in Section 2.1 except that the shift parameter, M, was changed from $M = 100$ to $M = 33$, and the analysis frame length, N, was changed from $N = 300$ to $N = 297$. Thus, for the long-time analysis, analysis frames were computed every 5 ms rather than every 15 ms, thereby giving a frame rate three times larger. The analysis frame was changed to 297 samples so as to be an integral multiple of M, the shift parameter. We denote the long-time LPC features as T_{LT}.

For convenience the short-time analysis was implemented with the same processing (i.e., that of Fig. 2) as that of the long-time analysis, except that N was changed to 99 (15-ms analysis frames) and M was again set to 33 (5-ms frame shifts). The order of the LPC analysis was kept at 8 for the short-time as well as the long-time analysis. We denote the short-time LPC features as T_{ST}.

To understand how the stationarity profile, μ, is generated within the framework of the LPC analysis, we must first define a characterization of the types of speech segments that are encountered. For this purpose we define two binary features that characterize the source and the dynamics of the vocal tract. The first feature describes the excitation for the frame of speech and we denote voiced speech as V, and unvoiced speech as $\bar{V}$. The second feature describes the vocal tract dynamics and we denote the stationary, steady-state case as SS, and the nonstationary, time-varying case as $\overline{SS}$. Thus, a given frame of speech is characterized by the notation $(V/\bar{V}, SS/\overline{SS})$.

The determination of whether a frame is voiced or unvoiced is fairly straightforward and is readily obtained from any number of pitch-detection algorithms. The determination of whether a frame is stationary or nonstationary is somewhat more complicated. This computation is made as follows. The basic idea is to compare both the long-time and short-time features of frames j and i, where j represents the frame occurring 15 ms before frame i. A distance comparing frames i and j is made as

$$\alpha_i = \frac{d[T_{LT}(i), T_{LT}(j)] + d[T_{LT}(j), T_{LT}(i)] + d[T_{ST}(i), T_{ST}(j)] + d[T_{ST}(j), T_{ST}(i)]}{4}, \tag{3}$$

i.e., the average of the long- and short-time LPC distances between

Table I—Feature sets used for similarity determination

Test Frame Status	Feature Set	Frame Spacing	Speech Example
(V, SS)	LT analysis	15 ms	Vowels, steady-state sounds
$(V, \overline{SS})$	LT analysis	5 ms	Onset, offset of voicing transitions
$(\overline{V}, SS)$	LT analysis	15 ms	Steady fricatives
$(\overline{V}, \overline{SS})$	ST analysis	5 ms	Transients

frames i and j and between frames j and i (recall that the LPC distance is not symmetric). The distance score, α_i, is then compared with a threshold (different for voiced and unvoiced frames), and the stationarity value is given as

$$SS = \begin{cases} 1 & \text{if } \underline{V} \text{ and } \alpha_i \leq \text{THV} \\ 1 & \text{if } \overline{V} \text{ and } \alpha_i \leq \text{THU} \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where 1 represents a stationary frame, and 0 represents a nonstationary frame, and THV and THU are voiced and unvoiced thresholds, respectively.

Once a frame has been characterized with the two-feature code, $(V/\overline{V}, SS/\overline{SS})$, the only remaining step is to specify which feature set and frame spacing should be used in the DTW distance computation.

It should be clear that for voiced frames, $(V, -)$, the long-time analysis should be used to avoid potential bias caused by the pitch period. Similarly, for all nonstationary frames, $(-, \overline{SS})$, a frame spacing of 5 ms should be used to track the fast dynamics of such frames. Finally, for unvoiced, nonstationary frames, $(\overline{V}, \overline{SS})$, the short-time analysis is most appropriate to follow transients and other brief events.

Table I shows a summary of the feature sets and frame spacings, for each of the four types of frames, as used to determine word and reference template similarity.

To illustrate the above analysis, Fig. 5 shows a series of plots of (a) the waveform, (b) the log energy (in dB), (c) the pitch, and (d) the average of long- and short-time LPC distance [eq. (3)] for the word /B/. It can be seen in Fig. 5a that the LPC distance becomes large at the beginning of voicing (point A in the plot), at the termination of voicing (point B in the plot), and at the end of the word (point C in the plot). Such frames (and their neighborhoods) are the nonstationary regions of the word, and generally correspond well with transients, onset and offset of voicing, and rapidly varying vocal-tract dynamics.

To determine the stationarity thresholds intelligently, THV and THU, histograms of the behavior of α_i for voiced and unvoiced frames, had to be measured. Such histograms are shown in Fig. 6. The data in this figure were obtained by computing α_i every 5 ms for all the frames of a 39-word vocabulary of letters of the alphabet plus the digits. Based

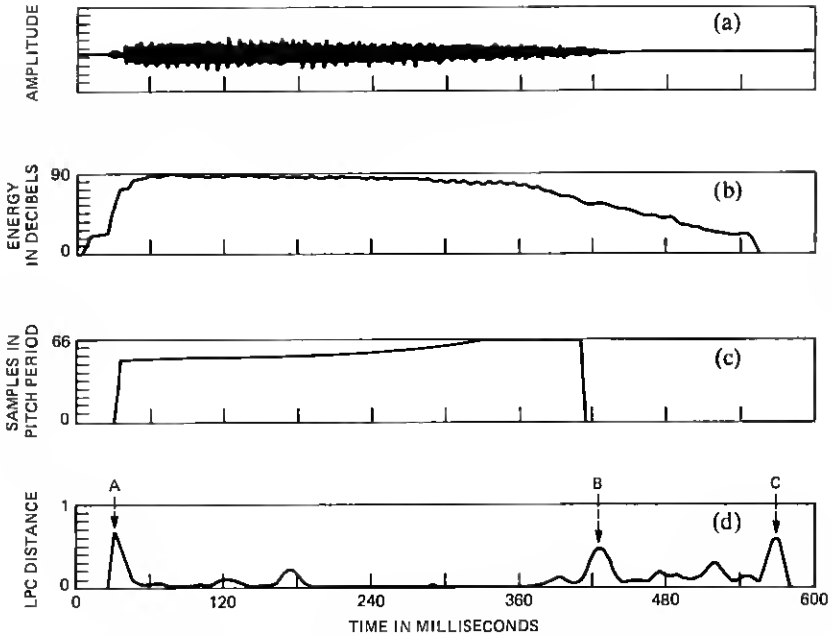


Fig. 5—An example (the word *B*) showing: (a) the waveform, (b) its energy profile, (c) its pitch contour, and (d) the LPC distance comparing adjacent frames.

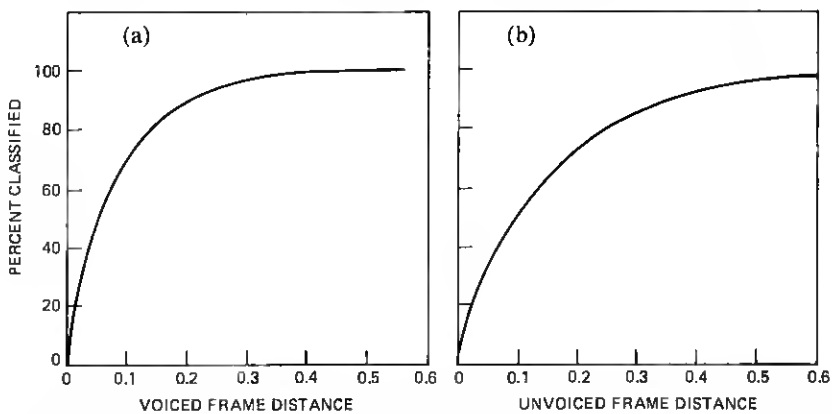


Fig. 6—Histograms of values of (a) LPC distance for voiced speech, and (b) unvoiced speech. Thresholds THV and THU are chosen to give desired percentages of nonstationary classification.

on the data of Fig. 6, values for THV and THU can be chosen, so as to obtain any desired average probabilities of occurrence of voiced or unvoiced classification. For example, if we assume that, on average, only 10 percent of the voiced frames should be classified as *SS*, then

a threshold of THV = 0.2 should be used. Similarly, for non-voiced frames a threshold of THU = 0.3 yields an average of 10 percent of the frames being classified as nonstationary. If the thresholds, THV and THU, are both set to infinity, then all frames are classified as stationary and hence the feature analysis is essentially identical to that of the original model. Similarly, if the thresholds are both set to zero, all frames are classified as nonstationary and a 5-ms frame spacing is used with both short- and long-time feature sets.

3.2 Modifications to the DTW algorithm for the improved word model

As discussed above, the basic changes made in the feature measurement were inclusion of both short- and long-time LPC analyses, and an increase in the frame rate of the analysis from once every 15 ms to once every 5 ms. These analysis changes required some modifications to the DTW algorithm to properly handle the raw data structure. The modifications primarily involve reformulation of the local path constraints to account for the diffferent possible frame spacings (i.e., nonuniform sampling in time), and modifications to the distance computation to handle both long- and short-time LPC distances and their appropriate weights.

We denote the long-time test pattern as $\{T_{LT}(n), n = 1, 2, \cdots, NT\}$, the short-time test pattern as $\{T_{ST}(n), n = 1, 2, \cdots, NT\}$, and the stationarity distance (on which the stationarity profile is based) as $\{\alpha_n, n = 1, 2, \cdots, NT\}$. Similarly, we denote the long-time reference pattern as $\{R_{LT}(m), m = 1, 2, \cdots, NR\}$ and the short-time reference pattern as $\{R_{ST}(m), m = 1, 2, \cdots, NR\}$.

We wish to solve for the optimum warping path of the form $m = w(n)$, defined for values of n that satisfy either of the following conditions:

$$(n - 1) \oplus 3 = 0 \tag{5a}$$

or

$$\alpha_n > \text{TH} \quad \text{or} \quad \alpha_{n-1} > \text{TH} \quad \text{or} \quad \alpha_{n-2} > \text{TH}. \tag{5b}$$

Equation (5a) says we solve for $m = w(n)$ at each standard 15-ms time slot. This constraint essentially guarantees a grid spacing, between adjacent DTW frames, of no more than three frames. It also guarantees that, in the limit, as the entire word is classified as stationary, the new analysis becomes identical to the previous analysis. Equation (5b) says we solve for $m = w(n)$ at each frame, n, in which the stationarity distance, α_n, of that frame or either of its two predecessors falls below the specified threshold, TH. (For voiced frames the threshold TH is set to THV, and for nonvoiced frames the threshold TH is set to THU). Cases in which eq. (5b) is satisfied (i.e., one of the distances is

above threshold) correspond to voiced frames with a rapidly changing spectrum (transitions), or unvoiced frames with nonstationary excitation.

For each frame n that satisfies one of the constraints of eq. (5) we must solve the DTW recursion

$$D_A(n, m) = w(n - n_L)\hat{d}(n, m) + \min_{\tilde{m}_L \le m_0 \le \tilde{m}_H} [D_a(n - n_L, m_0)], \qquad m_L \le m \le m_H, \quad (6)$$

where

n_L = last value of n for which a DTW recursion was done.
$\hat{n}_L$ = next-to-last value of n for which a DTW recursion was done.
$w(n - n_L)$ = weighting function on the local distance to account for the nonuniform frame spacing.
$\hat{d}(n, m)$ = local frame distance for reference frame m and test frame n.
$\tilde{m}_L$ = smallest value of m at $n = n_L$ from which a valid path can go to the grid point (n, m).
$\tilde{m}_H$ = largest value of m at $n = n_L$ from which a valid path can go to the grid point (n, m).
m_L = smallest value of m at frame n for which DTW recursion is solved.
m_H = largest value of m at frame n for which DTW recursion is solved.

The values of m_L and m_H are determined from the global path constraints which specify that all valid DTW paths must lie within a parallelogram defined from lines of slope 2 and slope 1/2 beginning at grid point (0, 0) and ending at grid point (NT, NR). Thus, m_L and m_H satisfy the path constraints

$$m_L = \max[(n - 1)/2 + 1, 2 \times (n - NT) + NR, 1] \quad (7a)$$

$$m_H = \min[2 \times (n - 1) + 1, (n - NT)/2 + NR, NR]. \quad (7b)$$

The values of $\tilde{m}_L$ and $\tilde{m}_H$ are those which guarantee that the path to grid point (n, m) satisfies the local constraint that the average slope be no less than one half nor more than 2. If we define a path increment function, $\Delta(m)$, as

$\Delta(m)$ = increment in m along the best path to grid point (n_L, m),

i.e., if the best path to grid point (n_L, m) comes from grid point $[\hat{n}_L, m - \Delta(m)]$, then values of m_0 in the DTW recursion [eq. (6)] must satisfy the local path constraint

$$\frac{(n - \hat{n}_L)}{2} \leq \Delta(m_0) + (m - m_0) \leq 2(n - \hat{n}_L). \tag{8}$$

Since $\Delta(m_0)$ also satisfies the constraint

$$\Delta(m_0) \leq 2\,(n_L - \hat{n}_L), \tag{9}$$

we can rewrite the inequalities of eq. (8) as

$$\tilde{m}_H - \Delta(\tilde{m}_H) \leq \frac{m - (n - \hat{n}_L)}{2} \tag{10a}$$

$$\tilde{m}_L \geq m - 2(n - n_L). \tag{10b}$$

Equation (10a) must be checked for each possible m value to find its solution, whereas eq. (10b) can be used directly.

The weighing function $w(n - n_L)$ is simply

$$w(n - n_L) = (n - n_L) \tag{11}$$

to give more weight to longer frame separations, and the distance $\hat{d}(n, m)$ of the form

$$\hat{d}(n, m) = \begin{cases} d[T_{LT}(n), R_{LT}(m)] & \text{if} \quad (V, SS), (\overline{V}, SS) \\ & \text{or} \quad (V, \overline{SS}) \qquad (12a) \\ d[T_{ST}(n), R_{ST}(m)] & \text{if} \quad (\overline{V}, \overline{SS}). \qquad (12b) \end{cases}$$

The complicated form of the DTW recursion is due to the nonuniform sampling rate at which the recursion is solved. If we translate eqs. (6) through (12) into words we can say that for each frame n for which the recursion is solved we compute $D_A(n, m)$ for a range of m from $m = m_L$ to $m = m_H$, as determined by the global path constraints. For each m the optimal path is determined as the weighted local distance, $\hat{d}(n, m)w(n - n_L)$, (as determined by the stationarity profile at frame n) plus the best accumulated distance to a predecessor frame that is a valid candidate for a path to frame m (i.e., $\tilde{m}_L \leq m_0 \leq \tilde{m}_H$). The range on m_0 is chosen to guarantee that the local path constraints of a warping curve slope of between 1/2 and 2 are met. Since the number of frames between the current frame n and the predecessor frame n_L, for which the DTW recursion was last solved, is variable (ranging from 1 to 3), the local path constraints must use this range, along with information as to how much the local path rose $[\Delta(m)]$ at frame (n_L, m_0) to set the local path constraints correctly.

The DTW recursion of eq. (6) is solved for all valid points from $n = 1$ to $n = NT$, and the total DTW solution is then given as

$$D^* = D_A(NT, NR) \tag{13}$$

and the average path distance is

$$\bar{D} = \frac{D_A(NT, NR)}{NT}. \tag{14}$$

3.3 The improved decision rule

As we discussed earlier a two-pass decision rule is used to improve recognition accuracy. The task of the first-pass decision rule is to determine the set of vocabulary words that are acoustically similar to the test word (i.e., the set of confusions). The task of the second-pass decision rule is then to resolve these confusions.

The key idea behind the operation of the second-pass decision rule is that the DTW distance scores between the test pattern and those reference patterns that are acoustically close to each other and to the test pattern consist of a χ^2 random component and a Gaussian random component. The χ^2 random component is associated with the averaging of distance scores between frames with the same basic spectrum, and therefore has a χ^2 distribution with p degrees of freedom. The Gaussian random difference is associated with the averaging of large distance scores between frames with dissimilar spectra.

In cases where the size of the dissimilar region is small (such as in comparing a /B/ to a /D/) compared to the size of the similar region, the χ^2 component distance often outweighs the Gaussian component, thereby causing potential recognition errors.

The purpose of the second-pass decision rule is to enhance the role of the Gaussian component associated with spectrally dissimilar regions in determining the final decision. This is accomplished using a distance-weighting function that enhances the discrimination power of the frame-by-frame distance scores.

By way of example, consider a simple confusion list of two references, R_i and R_j, for test word T. Let the DTW frame-by-frame distance and warping path be specified as

$$d_k(n) = d\{T(n), R_k[w(n)]\} \tag{15}$$

and

$w_k(n)$ = Warping path comparing frame n of the test with reference R_k.

We now define two distance-weighting functions,

$$\{W^{i,j}(n), n = 1, 2, \cdots, NR_i\}$$

$$\{W^{j,i}(n), n = 1, 2, \cdots, NR_j\},$$

where $W^{i,j}(n)$ is the weighting to discriminate R_j from R_i, and $W^{j,i}(n)$ is the weighting to discriminate R_i and R_j. (Reference 11 shows that

these weighting functions are generally not symmetric). We defer a discussion of how the weights are generated, in a speaker-trained system, to Section 3.4.

The basic hypothesis is that the test pattern, T, corresponds to either R_i or R_j, and we wish to come up with a discrimination score that aids in this decision. If we define a discrimination score, $\delta(T, R_i | T \in R_j)$, as the weighted distance between the T and R_i, assuming that T actually corresponds to R_j, then we get

$$\delta(T, R_i | T \in R_j) = \frac{\sum_{k=1}^{NT} W^{i,j}[w_i(k)] d\{T(k), R_i[w(k)]\}}{\sum_{k=1}^{NT} W^{i,j}[w_i(k)]}, \tag{16}$$

and similarly we get

$$\delta(T, R_j | T \in R_i) = \frac{\sum_{k=1}^{NT} W^{j,i}[w_j(k)] d\{T(k), R_j[w(k)]\}}{\sum_{k=1}^{NT} W^{i,j}[w_i(k)]}. \tag{17}$$

The weighted distance corresponding to the hypothesis $T \in R_j$ [i.e., eq. (16)] is shown in Fig. 7. The frame-by-frame distance is multiplied by the weighting function reflected through the warping curve to give the discrimination score δ.

The discrimination distances of eqs. (16) and (17) have the following important property. If T and R_i are from the same word (different replications) then the frame-by-frame distances, $d(\cdot, \cdot)$ are all χ^2 distributed (theoretically) and thus $\delta(T, R_i | T \in R_j)$ will be theoretically "independent" of the weighting function. If, however, T and R_j (instead of R_i) are from the same word, then $\delta(T, R_i | T \in R_j)$ will reflect to a greater extent the Gaussian-distributed component of the original distance score, $d(T, R_i)$, since it primarily consists of distance in regions where R_j and R_i differ significantly, even though they may be quite short.

Thus, in the simple case of a confusion between two references, R_i and R_j, the final decision is made on the basis of the discrimination scores of eqs. (16) and (17).

More generally, if the confusion list associated with test pattern T has Q candidates, $\{R_{i_1}, R_{i_2}, \cdots, R_{i_Q}\}$, then the following procedure is followed:

(i) Compute all pairs of discriminations

$$\delta(T, R_{i_a} | T \in R_{i_b}), b \neq a, a, b = 1, 2, \cdots, Q\}.$$

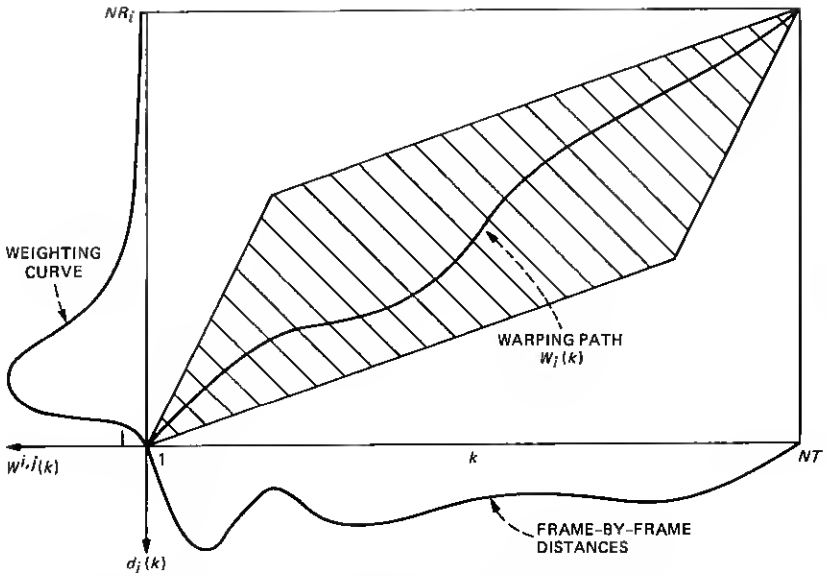


Fig. 7—The time warping plane of a test and reference pattern along with the distance of each frame and the weighting curve on distance.

(*ii*) Form the average discrimination distance

$$\bar{\delta}(T, R_{i_a}) = \frac{1}{Q-1} \sum_{\substack{b=1 \\ b \neq a}}^{Q} \delta(T, R_{i_a} | T = R_{i_b}), \; a = 1, 2, \cdots, Q.$$

(*iii*) Define the most likely candidate, R_i, as the candidate with the minimum average discrimination distance, i.e.,

$$\delta_{\text{MIN}} = \min_a \{\bar{\delta}(T, R_{i_a})\}.$$

Similarly, a least likely candidate with maximum distance is defined as

$$\delta_{\text{MAX}} = \max_a \{\bar{\delta}(T, R_{i_a})\}$$

(*iv*) Given the original (i.e., first-pass) distance scores for all Q candidates, $d(T, R_{i_a})$, with smallest distance d_{MIN} and largest distance d_{MAX}, a second-pass set of distances scores is computed by retaining second-pass ordering with first-pass distances. This procedure is illustrated in Fig. 8. A reference with second-pass discrimination score $\bar{\delta}(T, R_i)$ is given distance $\bar{d}(T, R_i)$ by linearly interpolating along the line of Fig. 8.

3.4 Determination of the weighting curves in the speaker-trained case

The determination of the weighting curves, $W^{j,i}$ and $W^{i,j}$, is readily

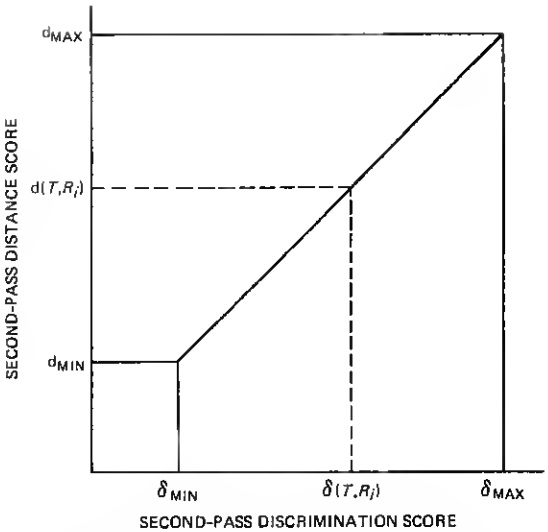


Fig. 8—Linear transformation between second-pass distance score and second-pass discrimination score.

performed in the training phase for speaker-trained systems. Given reference templates R_i and R_j, as obtained using the robust training procedure of Rabiner and Wilpon,[19] a simple way of obtaining $W^{i,j}$ is to warp R_i to R_j, giving

$$W^{i,j}(k) = d\{R_j(n), R_i[w_j(n)]\}, \tag{18}$$

where $w_j(n)$ denotes the warping path. Thus, the frame weights (W) are essentially the frame-by-frame warped DTW distances between the reference templates. Figure 9 shows weighting functions for references corresponding to the words /*I*/ and /*Y*/. When compared with the speaker-independent weights of Rabiner and Wilpon,[11] we immediately see the statistical effects of small samples. It is evidence that the curves of Fig. 9 need some smoothing to reduce the statistical variance. The resulting of applying a 3-point smoother (a triangular window) to the data of Fig. 9 is given in Fig. 10. A good deal of the statistical variation in the curves is smoothed out.

An alternative, more statistically meaningful, way of obtaining smoother weighting curves is to use all P replications of each word in the training set to determine the weights. Basically, we obtain a weighting function for each pair of training tokens such that each token is close in distance to the appropriate reference. The final weighting curve is then obtained by averaging the individual weighting curves, with appropriate time alignments. We use the term subweights to denote the set of weights obtained by averaging all training tokens,

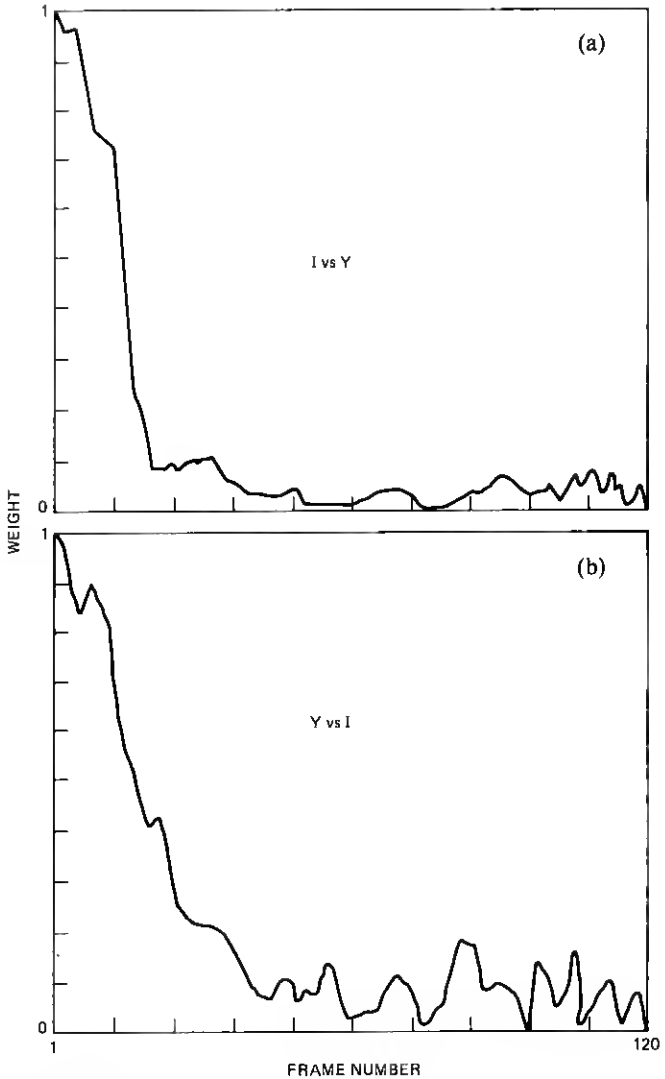


Fig. 9—Typical weighting curves for (a) I vs Y, and (b) Y vs I, derived from the robust training tokens.

and we use the notation S to refer to this set. Figure 11 illustrates the (sub) weighting curves for I, Y comparisons based on a set of five training tokens for each word.

IV. EXPERIMENTAL EVALUATION OF THE IMPROVED MODEL

To measure the performance of the improved, LPC-based, isolated word-recognition model, a small evaluation test was performed. Each

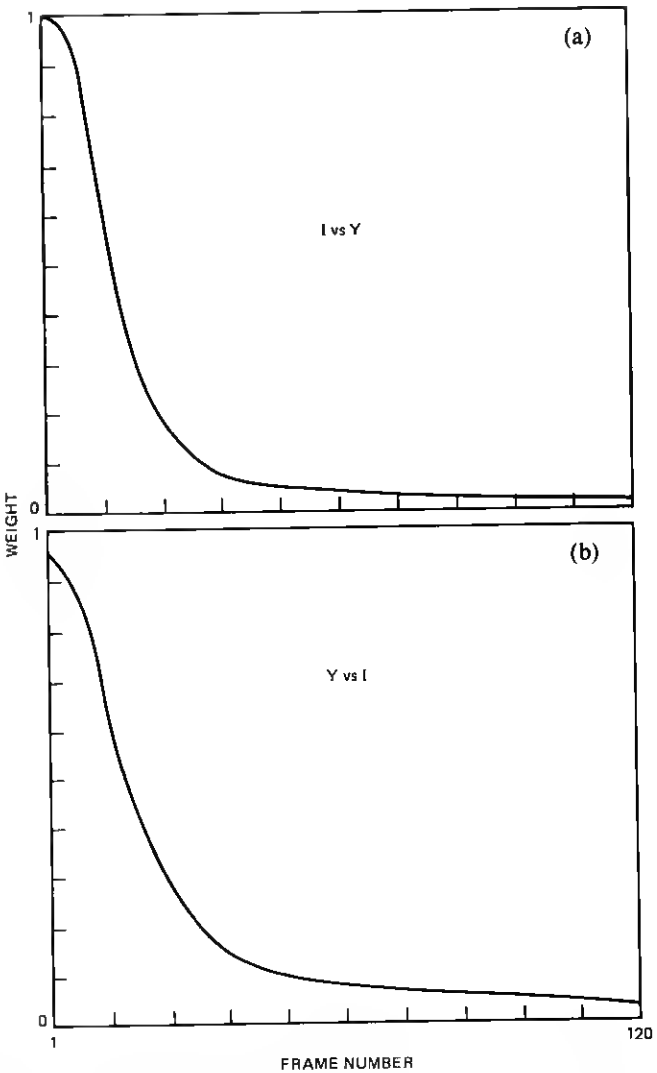


Fig. 10—Smoothed weighting curves for (a) I vs Y, and (b) Y vs I, derived from the robust training tokens and a 3-point smoother.

of four talkers (two male, two female—all experienced with speech-recognition systems) trained the recognizer on a 39-word alpha-digit vocabulary by saying each vocabulary word five times during the course of a single training session. The word-reference patterns, the normal discrimination weights, W, and subweights, S, were determined from the training data using the robust training procedure of Rabiner and Wilpon.[19]

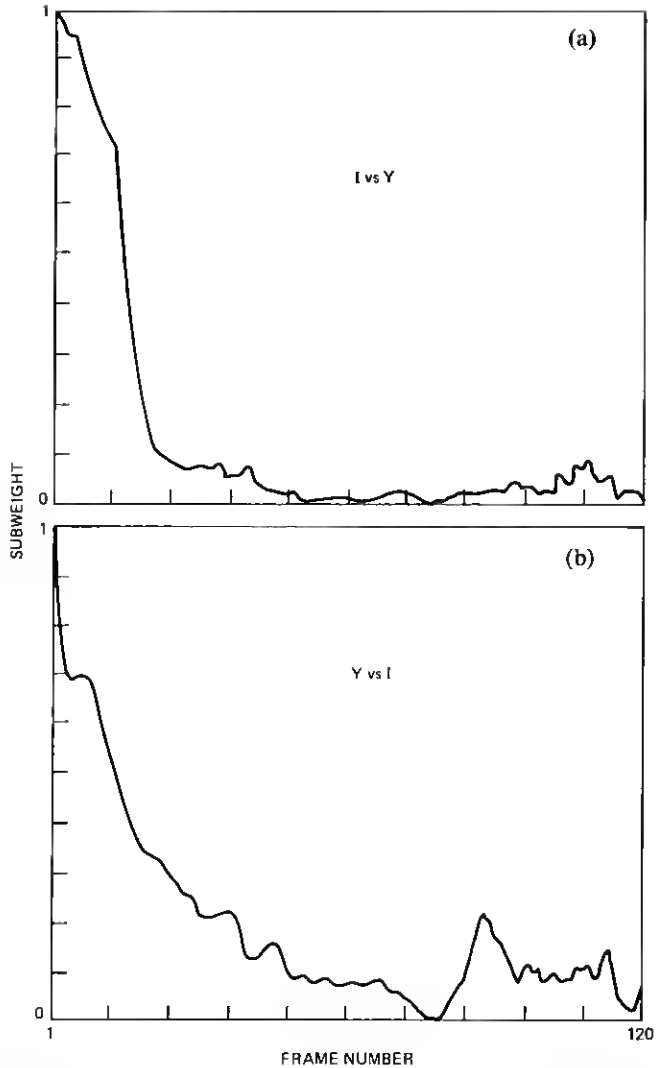


Fig. 11—Subweight curves for (a) I vs Y, and (b) Y vs I, derived from using all training tokens.

For evaluation purposes the 39-word vocabulary was spoken 10 additional times by each of the four talkers in two distinct recording sessions. Thus, a total of 390 words were used in each recognition test for each talker.

4.1 Recognition test results

The overall results of the evaluation tests are given in Table II,

Table II—Recognition accuracies as a function of the stationarity thresholds and the number of recognition passes for the four talkers

		Talker Number			
	(THU, THV)	1	2	3	4
Pass 1 Alone	$(-\infty, \infty)$	94.9	94.9	90.5	86.7
	(−.3, .2)	95.4	94.9	91.8	86.4
	(0., 0)	94.9	94.9	91.5	85.4
Pass 2 With Weight W	$(-\infty, \infty)$	96.7	95.6	94.1	87.2
	(−.3, .2)	96.4	95.6	94.9	88.5
	(0., 0)	95.6	95.4	94.4	86.4
Pass 2 With Subweight S	$(-\infty, \infty)$	95.6	95.9	95.4	87.2
	(−.3, .2)	95.4	95.4	96.2	87.9
	(0., 0)	95.4	95.6	95.1	87.2

which shows recognition accuracy as a function of stationarity thresholds, talker, and analysis condition. Three analysis conditions are shown, namely Pass 1 alone (no discriminant analysis), Pass 2 with weights, W, derived from single reference tokens, and Pass 2 with subweights, S, derived from all reference tokens.

The results of using Pass 1 alone show only a 0.4-percent improvement, on average, in recognition accuracy for the four talkers when comparing the old stationary model (where THU = $-\infty$, THV = ∞) with the new stationary model (where THU = −0.3, THV = 0.2).

The results of using Pass 2 with weights W show an average of 2.1-percent improvement in recognition accuracy for the four talkers over the old stationary model (when THU = −0.3 and THV = 0.2). When subweights S are used in Pass 2, the improvement in recognition accuracy is an average of 2 percent.

Table II also shows that when Pass 2 is used the recognition accuracy with stationarity thresholds set to (−0.3, 0.2) is, on average, about 0.5 percent higher than with stationarity thresholds set to $(-\infty, \infty)$. This result indicates that the improved model provides a consistent recognition accuracy improvement of about 0.5 percent, with or without the second-pass weights.

V. DISCUSSION

The results presented in Section IV are both encouraging and discouraging. They are encouraging in that real improvements in recognition accuracy were obtained when a nonstationary analysis framework was used in place of the purely stationary framework used in earlier work. They are discouraging in that the average improvement resulting from the nonstationary model (0.5 percent) was considerably smaller than the average improvement resulting from the discrimination analysis of the second pass (1.6 percent).

There are several points worth noting that have bearing on the discussion and results of this paper. The first concerns the anticipated improvement in performance for the improved word-recognition model. If one carefully considers the sources for recognition errors with the alpha-digit vocabulary, it should become clear that the anticipated improvement resulting from the nonstationary analysis should be small unless some extra weighting is applied to the nonstationary regions. This is because words that are strongly affected by the nonstationary analysis (e.g., *p*, *d*, *t*, *k*, etc.) are easily confused with similar words in the vocabulary (e.g., *b*, *v*, *g*, *a*, etc.), and since the nonstationary regions are only a small subset of the word patterns, the improved analysis will be swamped out by the word-similarity regions. This is the original motivation for the discriminant analysis model used in the two-pass word recognizer.[11] Hence, the results of Section IV, which show a small (but consistent gain) for the improved analysis model and a somewhat larger gain for the discriminant model, are entirely consistent with the anticipated results given above.

A second point of note is that the implementation of the improved word model was more of a convenient one, rather than one that naturally followed from the theory. Thus, the short-time features were LPC coefficient sets derived from a short-time window. This implementation was straightforward and required only minimal modification of the recognizer structure. A more reasonable implementation of the short-time analysis in the model would have been something like a filter bank model, or a basilar membrane model. Such features would then have complemented the long-time LPC features and would have provided a better vehicle for testing and evaluating the improved model. The problem with using these alternative short-time feature sets is that there is no simple way of combining LPC and filter bank (or basilar membrane model) features and deriving from them a distance measure with good physical properties. The problem of combining LPC and energy features has already been investigated by Brown and Rabiner,[20] and it was shown that no simple metric existed even for such a simple case. The main point in the above discussion is that the small gain of the improved word model is more impressive when one considers the simplicity of the short-time analysis used to provide the performance gain.

The third point of note is the fact that the simple weighting derived from the robust training procedure seemed to provide the same performance improvement as the more sophisticated weighting obtained by using multiple tokens in obtaining the weights. The obvious conclusion to be drawn from the result is that the gain obtained from the second pass (which is due primarily to small regions of extreme spectral difference) is manifested in any pair of training tokens and that simple

smoothing (to eliminate statistical variability) is as good as using multiple tokens.

When one takes into consideration all of the above points, the results of Section IV provide a reasonable basis for believing that the improved word-recognition model is a reasonable one and that both the nonstationary analysis of the first pass, and the discrimination analysis of the second pass provide real performance gains.

VI. SUMMARY

An improved word-recognition model was proposed in which the standard long-time analysis features of the model are combined with a set of short-time analysis features. A stationarity index is also computed for each speech frame indicating which set of features (long-time or short-time) best characterized the current frame of speech. Appropriate modifications to the DTW algorithm were required to handle the enhanced analysis feature set. Also incorporated in the recognition model was a speaker-trained version of the discriminant analysis, two-pass model proposed by Rabiner and Wilpon.[11]

An evaluation of the model based on an LPC implementation of both long-time and short-time feature sets showed the overall improved word model had from 1- to 5.7-percent improvement in recognition accuracy across four experienced users of speech recognition systems using an alpha-digit word vocabulary. On an average the nonstationary feature set alone led to a 0.5-percent improvement in accuracy, whereas the two-pass discriminant analysis alone led to a 1.6-percent average improvement in accuracy. The two improvements were almost independent and the overall recognizer had, on average, a 2.1-percent improvement in word accuracy.

The above results are considered encouraging and indicate that the improved model should be considered with alternative short-time feature sets.

REFERENCES

1. T. B. Martin, "Practical Applications of Voice Input to Machines," Proc. IEEE, *67* (April 1976), pp. 487–501.
2. N. R. Dixon and T. B. Martin, eds., *Automatic Speech and Speaker Recognition*, New York: IEEE Press, 1979.
3. W. Lea, ed., *Trends in Speech Recognition*, Englewood Cliffs, NJ: Prentice-Hall Inc., 1980.
4. D. R. Reddy, ed., *Speech Recognition*, New York: Academic Press, 1974.
5. L. R. Rabiner and S. E. Levinson, "Isolated and Connected Word Recognition—Theory and Selected Applications," IEEE Trans. on Commun., *COM-29*, No. 5 (May 1981), pp. 621–59.
6. G. M. White and R. B. Neely, "Speech Recognition Experiments with Linear Prediction, Bandpass Filtering, and Dynamic Programming," IEEE Trans. on Acoustics, Speech, and Signal Processing, *ASSP-24* (April 1976), pp. 183–8.
7. L. R. Rabiner, A. E. Rosenberg, J. G. Wilpon, and W. J. Keilin, "Isolated Word Recognition for Large Vocabularies," B.S.T.J., *61*, No. 10 (December 1982).

8. B. Aldefeld, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilpon, "Automated Directory Listing Retrieval System Based on Isolated Word Recognition," Proc. IEEE, *68* (November 1980), pp. 1364–79.
9. S. E. Levinson, "The Effects of Syntactic Analysis on Word Recognition Accuracy," B.S.T.J., *57* (May–June 1977), pp. 1627–44.
10. L. R. Rabiner, J. G. Wilpon, and A. E. Rosenberg, "A Voice-Controlled Repertory-Dialer System," B.S.T.J., *59* (September 1980), pp. 1153–63.
11. L. R. Rabiner and J. G. Wilpon, "A Two-Pass Pattern-Recognition Approach to Isolated Word Recognition," B.S.T.J., *60,* No. 5 (May–June 1981), pp. 739–66.
12. F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. Acoustics, Speech, and Signal Processing, *ASSP-23,* No. 1 (February 1975), pp. 67–72.
13. J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech,* New York: Springer-Verlag, 1976.
14. P. V. de Souza, "Statistical Tests and Distance Measures for LPC Coefficients," IEEE Trans. on Acoustics, Speech, and Signal Processing, *ASSP-25* (December 1977), pp. 554–9.
15. J. M. Tribolet, L. R. Rabiner, and M. M. Sondhi, "Statistical Properties of an LPC Distance Measure," IEEE Trans. on Acoustics, Speech, and Signal Processing, *ASSP-27,* No. 5 (October 1979), pp. 550–8.
16. H. Sakoe and S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," IEEE Trans. on Acoustics, Speech, and Signal Processing, *ASSP-26* (February 1978), pp. 43–9.
17. L. R. Rabiner, A. E. Rosenberg, and S. E. Levinson, "Considerations in Dynamic Time Warping for Discrete Word Recognition," IEEE Trans. on Acoustics, Speech, and Signal Processing, *ASSP-26* (December 1978), pp. 575–82.
18. C. S. Myers, L. R. Rabiner, and A. E. Rosenberg, "Performance Trade-offs in Dynamic Time Warping Algorithms for Isolated Word Recognition," IEEE Trans. on Acoustics, Speech, and Signal Processing, *ASSP-28* (December 1980), pp. 622–35.
19. L. R. Rabiner and J. G. Wilpon, "A Simplified, Robust Training Procedure for Speaker Trained, Isolated Word Recognition Systems," J. Acoust. Soc. Amer., *68,* No. 5 (November 1980), pp. 1271–6.
20. M. K. Brown and L. R. Rabiner, "On the Use of Energy in LPC Based Recognition of Isolated Words," B.S.T.J., *61,* No. 10 (December 1982).